DISCS

Jim Gray

Feb 1989

OUTLINE

DEBIT CREDT STANDARDIZATION

DISC TRENDS & ECONOMICS

DISC PHYSICS

DISC SUBSYSTEMS

Debit Credit Council

Renamed:

Transaction Processing Performance Council

Benchmark: TPC Benchmark A™

Members:

ATT,Biin, CDC,Computer Associates, Cullinet, DG, DEC, Fujitsu, HP, HB, IBM, ICL, Informix, NCR, Oracle, Prime, Pyramid, RTI, Sequent, Sequoia, Stratus, Sun, Sybase, Tandem, Teradata, Tolerant, Unisys, Wang

Harder:

Measure response time at driver system

Reply must return new balance

Easier

Shrink terminal net by 10X

Eliminate Presentation Services

Shrink history file by 3x

Response time: 90% @ 2 seconds (vs 95% @ 1 sec)

BIG DEBATE:

How to characterize the network?

LAN?

WAN?

Contact: Omri Serlin,
ITOM International
POB 1450
Los Altos, CA 94022
415-948-4516

OUTLINE

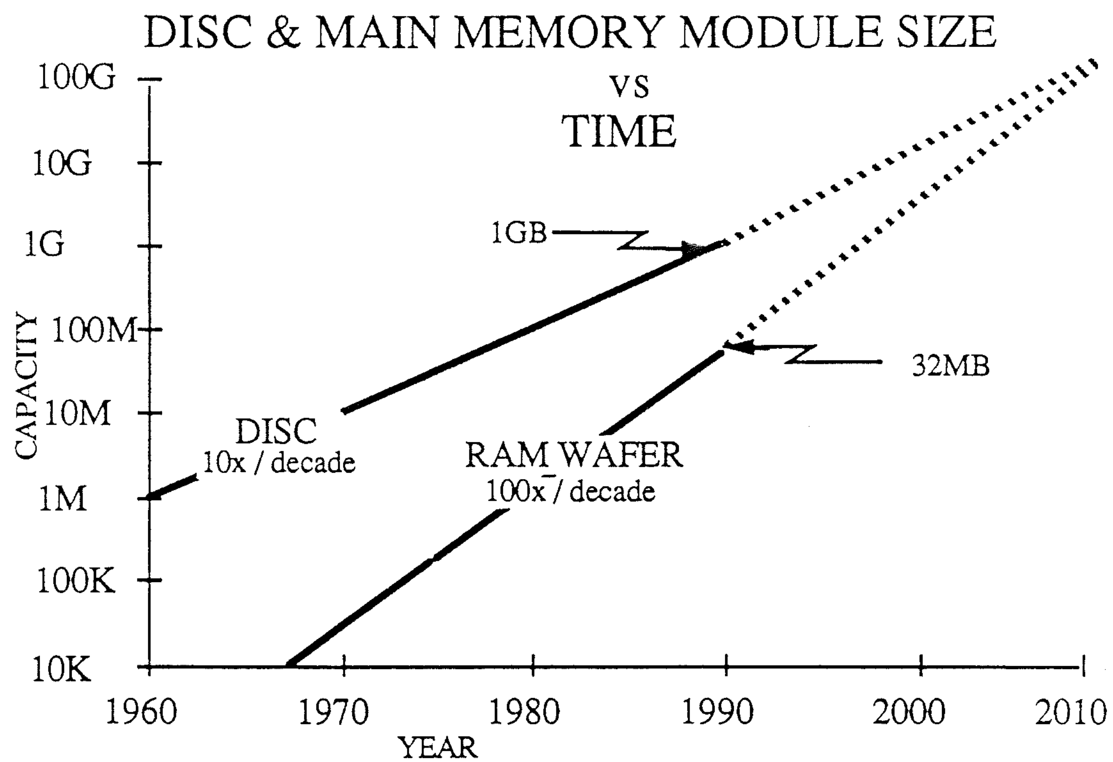
DEBIT CREDT STANDARDIZATION

DISC TRENDS & ECONOMICS

DISC PHYSICS

DISC SUBSYSTEMS

DISC ECONOMICS / TRENDS



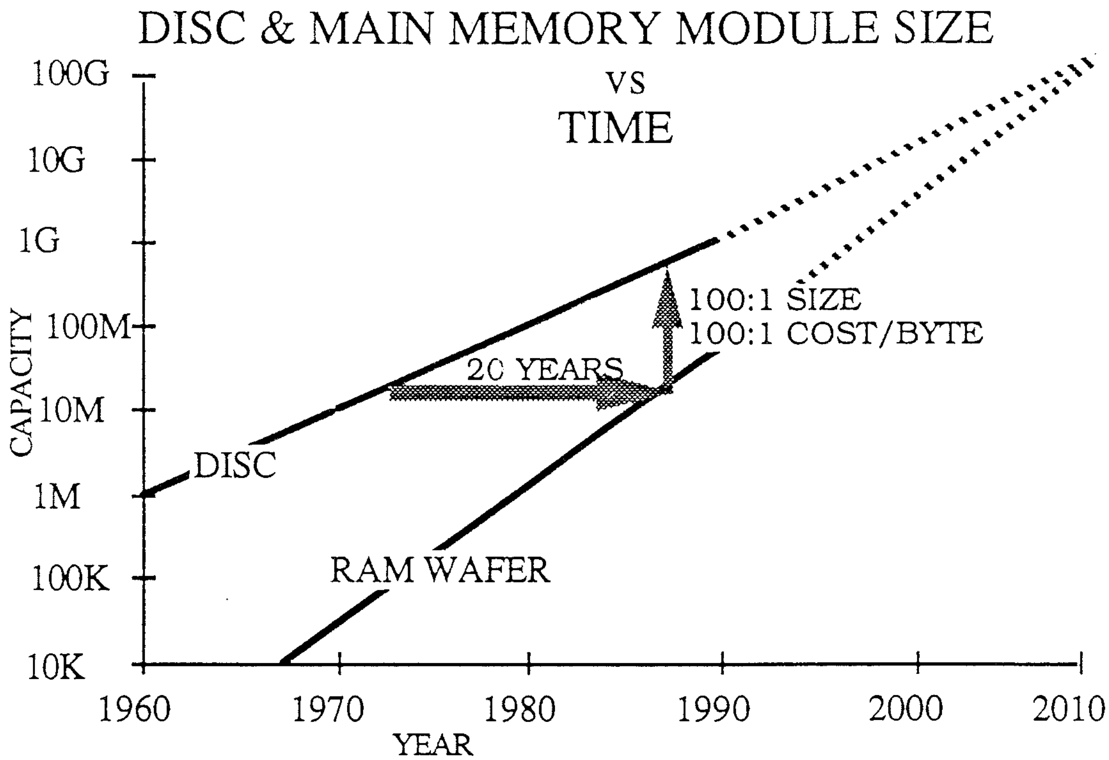
Hoagland: Disc Magnetic Areal Density (MAD) $= 10^{(year-1970)/10}$ Mb/in^2

Moore: RAM Memory Density $= 10^{(year-1970)/5}$ Kb/chip

Disc	~ 5$/MB- 20$/MB	.1$/access - 4k$/access
RAM:	100$/MB-5k$/MB	???????

Next Decade: Disc & Controller ~ 100$ ~1GB => .1$/MB
RAM Wafer: ~1K$ ~.5GB => 1$/MB

DISC ECONONOMICS TODAY



Someday: Disc will be "tape"

Cheap archive sequential storage,

NOT Random Block Access Storage Device

Today: 5 minute rule applies:
keep it in ram if accessed every 5 minutes

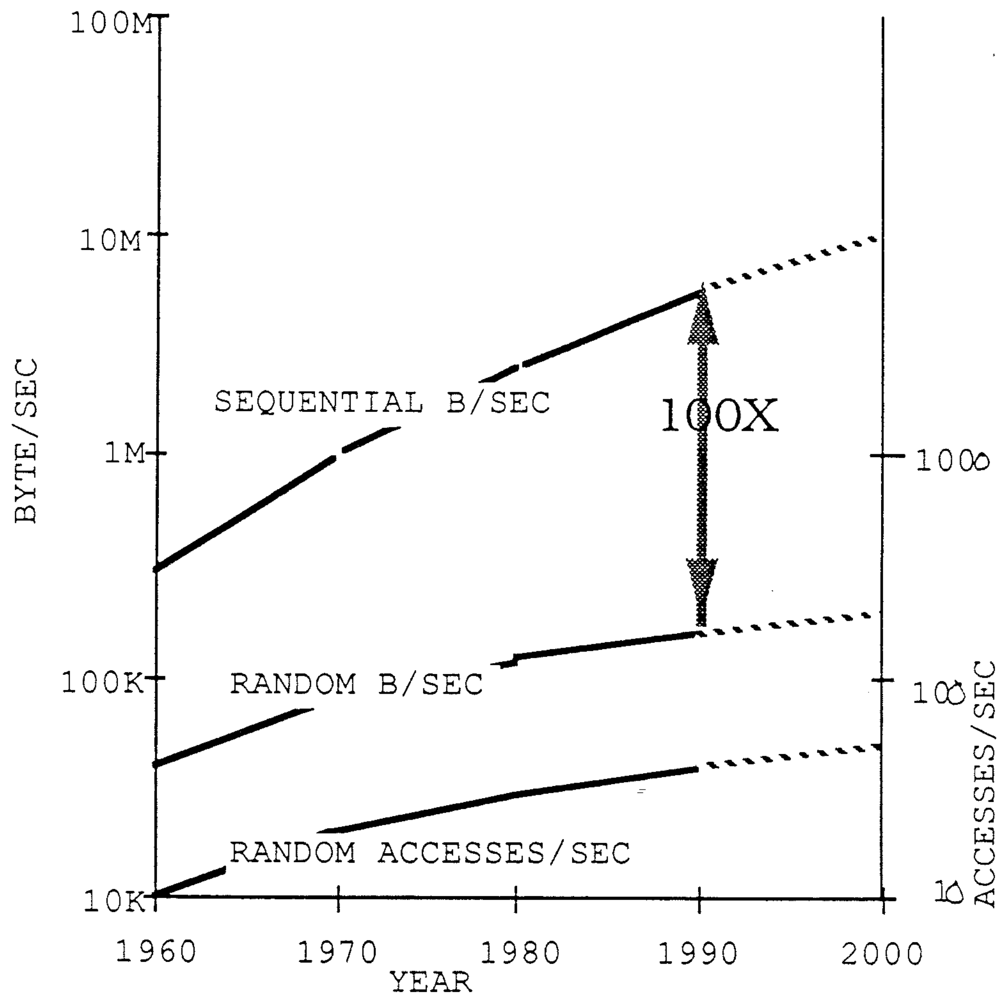
J. Gray , F. Putzolu, *The 5 Minute Rule for Trading Memory for Disc Accesses, and the 10 Byte Rule for Trading Memory for CPU Instructions*, ACM SIGMOD Proceedings, June 1987,

THE BIG DISC PROBLEM: Disc Delivers 25accesses/second:

100MB	1 a/s/4MB,
1GB	1 a/s/40MB
100GB	1 a/s/4GB

EVEN TODAY, DISC NEEDS TO BE USED SEQUENTIALLY

DISC SPEED vs TIME



1. ACCESS RATE NOT MUCH IMPROVED
2. SEQUENTIAL **100X** RANDOM

SO: USE SEQUENTIAL "DISC IS TAPE!"

LARGE BLOCK TRANSFERS

CONVERT RANDOM IO TO LOG IO

OUTLINE

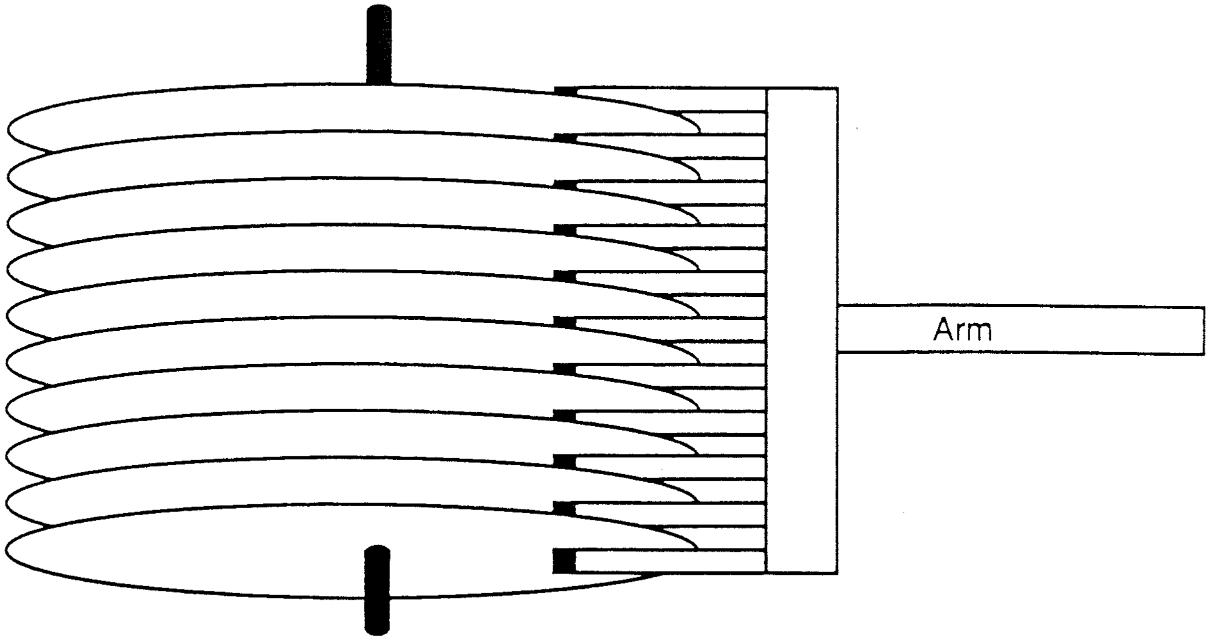
DEBIT CREDT STANDARDIZATION

DISC TRENDS & ECONOMICS

DISC PHYSICS

DISC SUBSYSTEMS

Laws of Nature



Discs rotate at 60rps (1800 -> 2400 -> 3600)

=> 60 io/sec max (50 due to creep)

=> ~16ms/rotation

May rise in future

Service_time = Seek + Settle + Rotate + Transfer

Settle ~ 2ms

Rotate ~ 1/2 (16ms) ~ 8ms

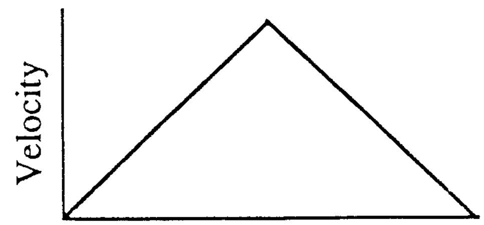
Work on Seek & Transfer

SEEK TIME

Seek_time ~ $\sqrt{\text{distance}}$

because: 1: constant acceleration

Velocity vs Time @ Constant Acceleration



2. area under curve (distance) ~ time^2

Expected seek distance:

If random access, then $\frac{1}{3}$ of total tracks

(difference of two random variables).

Trends:

As discs get smaller 14"-> 9"->8"->5 $\frac{1}{4}$"->3 $\frac{1}{2}$":

seek distance decreases (linear)

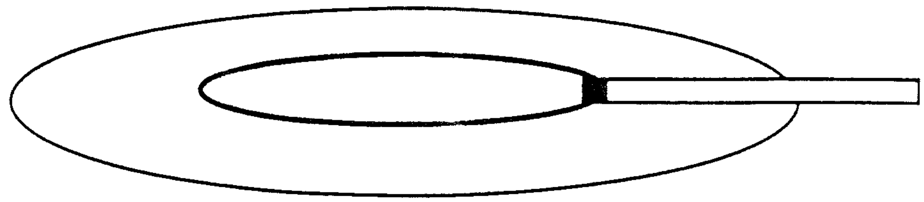
seek time decresases $\sqrt[2]{\text{stroke}}$

arms are $\sqrt[3]{\text{lighter}}$ => faster acceletation

less power, stress => reliable and cheap

MAD decrease implies less seek needed: $\sqrt[2]{10}$ ~ 3x/decade

TRANSFER TIME



Transfer_time ~ bytes/bandwidth

Typical Bandwidth: 1MB/s ... 10MB/s

Bandwidth ~ Rotations/sec * Bytes/track

but Rotations/sec ~ 60 is a universal constant so

~ Bytes/track

~ (Bytes/inch) * (inches/track)

~ $\sqrt{MAD}$ * Diameter

Trend: Discs are shirinking 14"-> 9"->8"->5$\frac{1}{4}$"->3$\frac{1}{2}$":

=> Diameter is shrinking (3x in this decade)

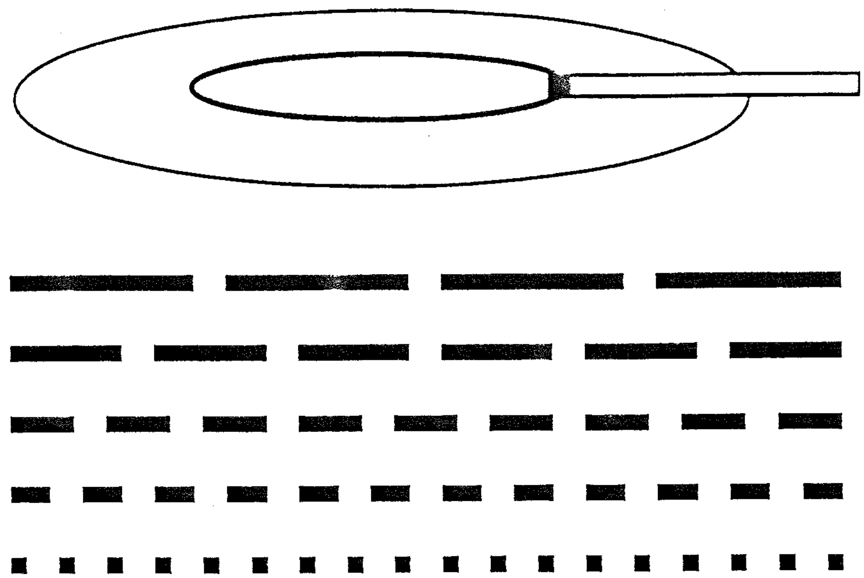
Perhaps this will end.

=> $\sqrt{MAD}$ decreases ~3/decade

Net: zero change in bandwidth

"Solution": Parallel read from multiple heads

FORMATTING



Disc Track formatted into **Blocks** or **Sectors** (512 is typical)

Separated by **Gaps**

Gaps are fixed by switching times,

speed of light to controler/cpu

As density increases, gaps dominate space.

At present 25% gap, 75% data is typical.

=> Formatted capacity ~ .75 rated capacity

=> Data Bandwidth ~.75 rated bandwidth

"**Solution**": Bigger blocks 4KB => 8x fewer blocks

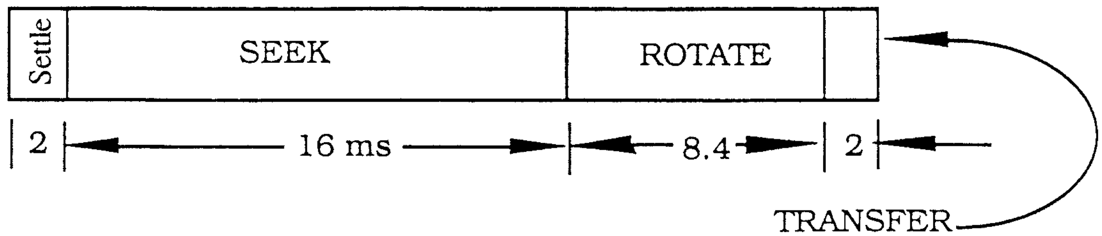
97% used space

At present 10µs ≈ 30 bytes.

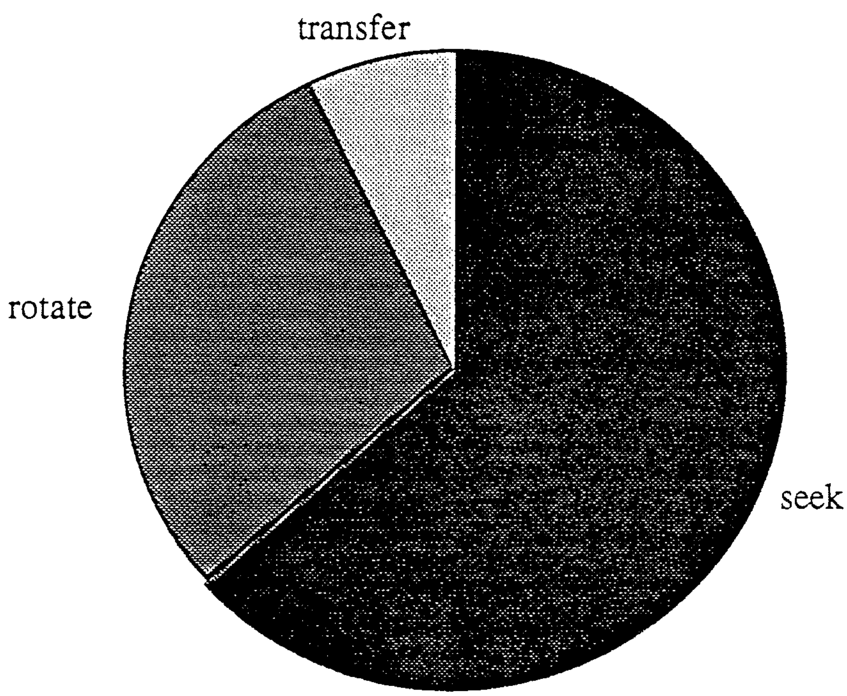
* Each block has address, checksum, etc

SUMMARY OF DISC PHYSICS

Service_time = Seek + Settle + Rotate + Transfer



Work on Queue, Seek & Transfer



OUTLINE

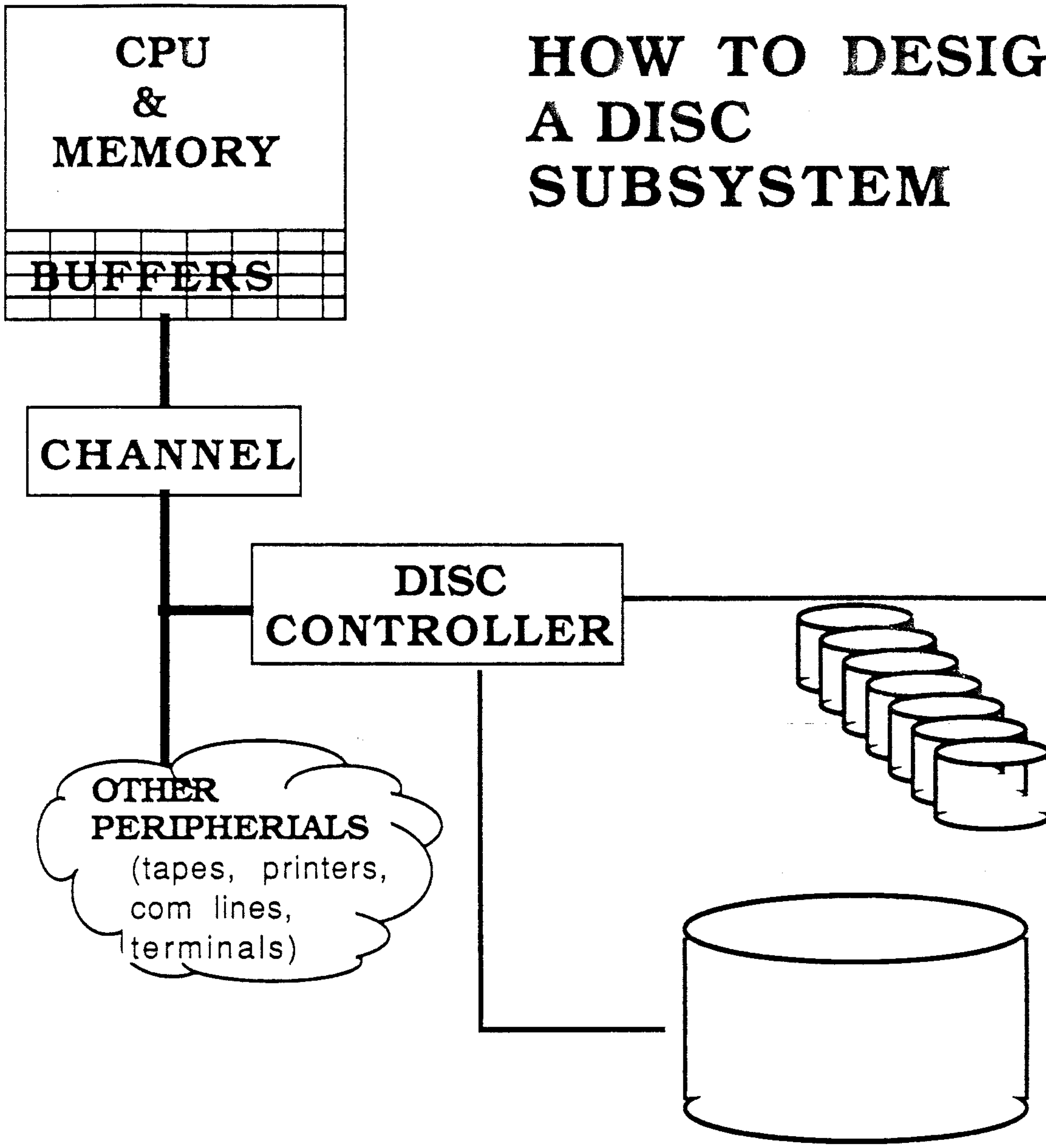
DEBIT CREDT STANDARDIZATION

DISC TRENDS & ECONOMICS

DISC PHYSICS

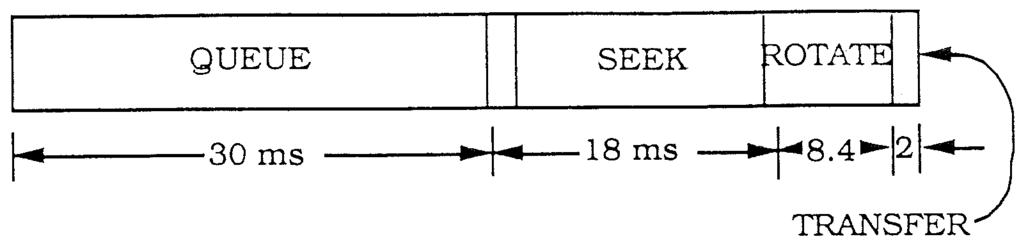
DISC SUBSYSTEMS

HOW TO DESIGN A DISC SUBSYSTEM

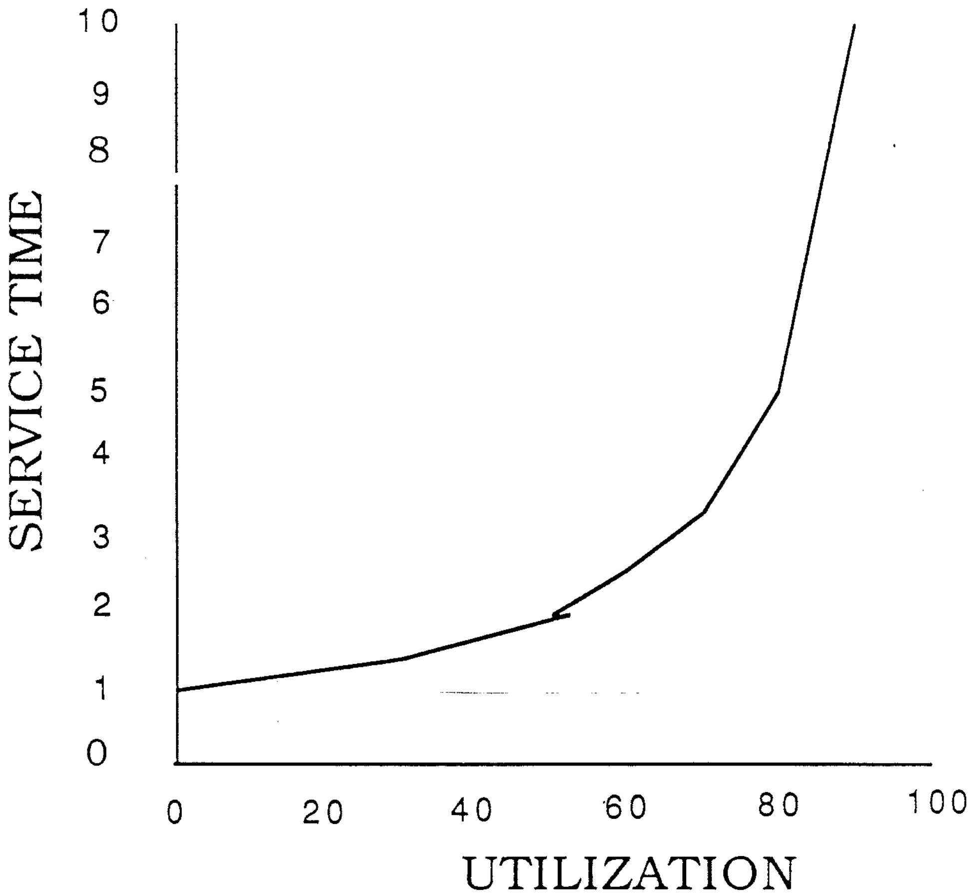


WHERE DOES THE ACCESS TIME GO?

PREDICTED:

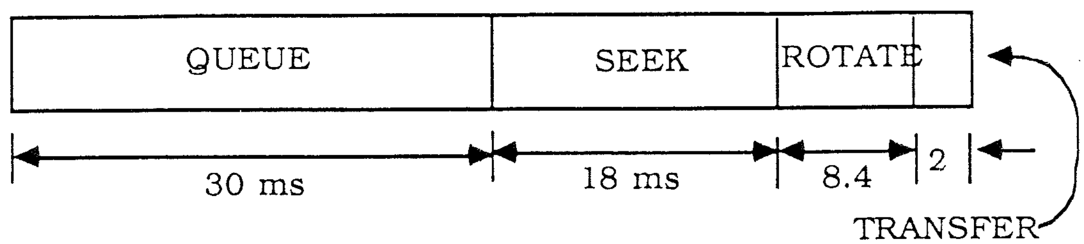


SERVICE TIME VS UTILIZATION

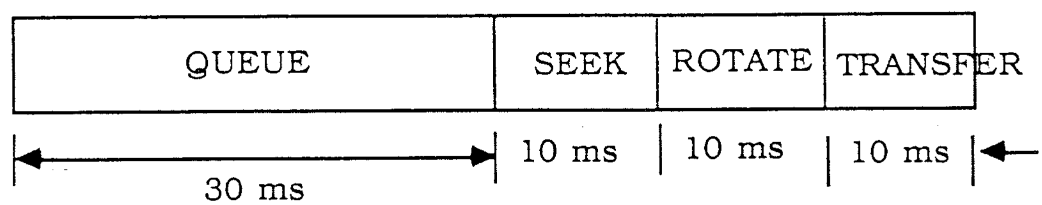


WHERE DOES THE ACCESS TIME GO

PREDICTED:



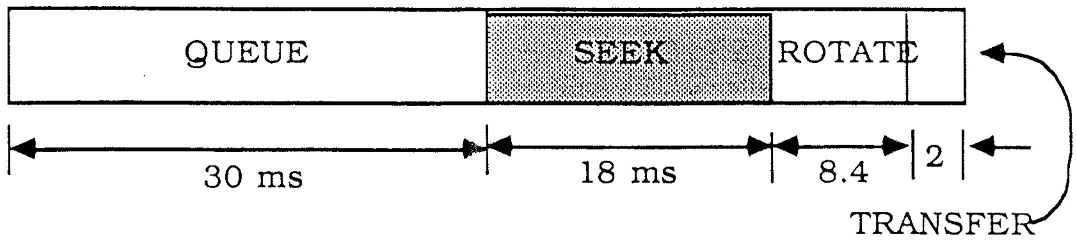
MEASURED:



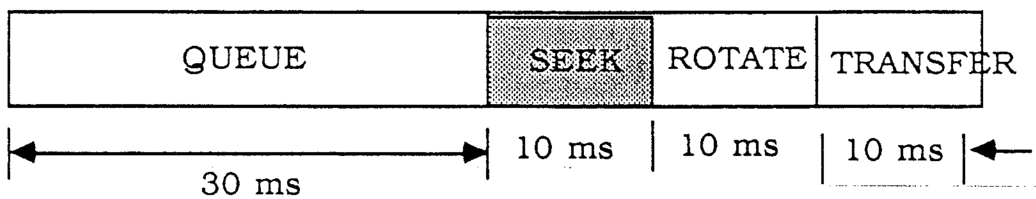
R.A. Scranton & D.A. Thompson, The Access Time Myth, IBM Research Report RC 10197 (#45223) 9/21/83

WHERE DOES THE ACCESS TIME GO

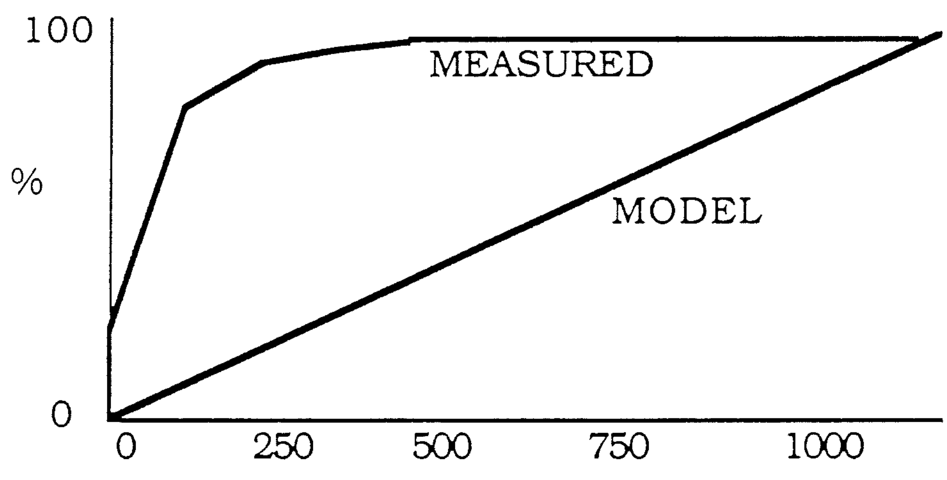
PREDICTED:



MEASURED:



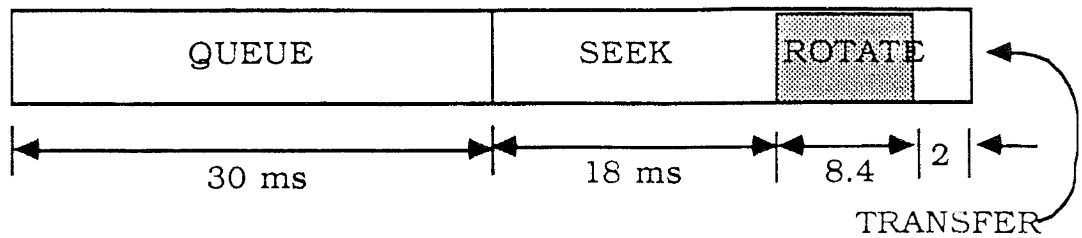
MOST SEEKS ARE SHORT



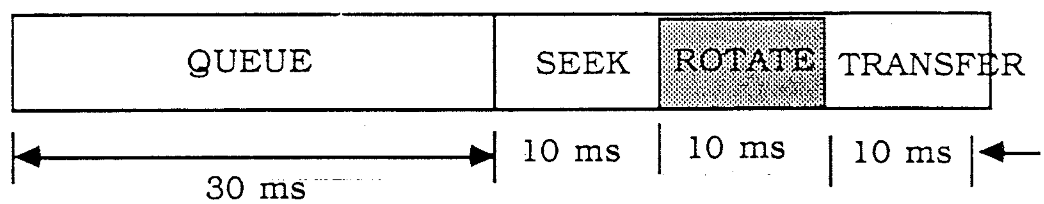
SUGGESTION: AVOID ZERO-LENGTH SEEKS

WHERE DOES THE ACCESS TIME GO

PREDICTED:



MEASURED:



10% RPS MISS BECAUSE

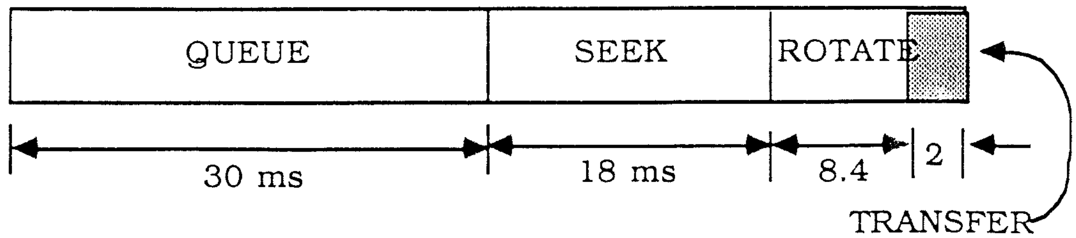
- CONTROLLER BUSY
- CHANNEL BUSY
- CPU BUSY

SUGGESTION:

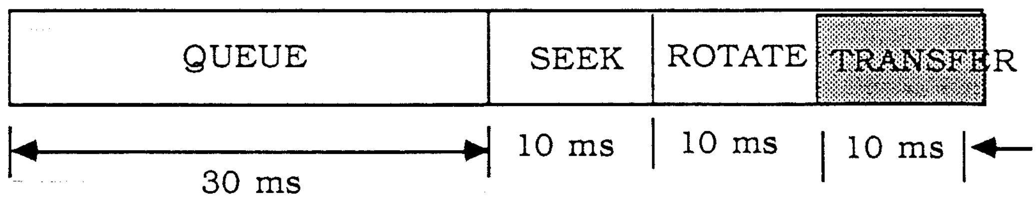
- ABANDON RPS
- PUT BUFFER ON DISC CONTROLLER

WHERE DOES THE ACCESS TIME GO

PREDICTED:



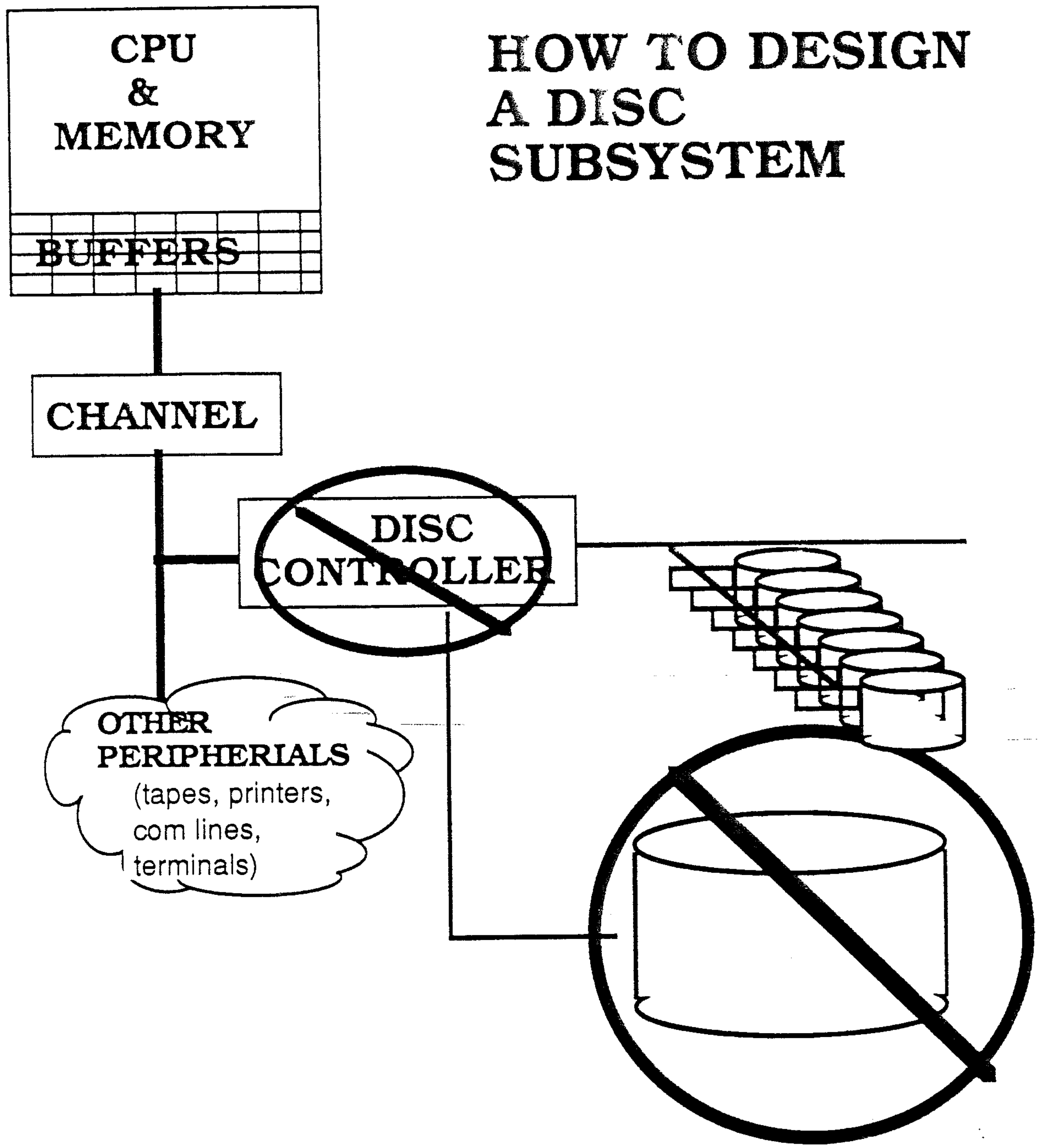
MEASURED:



CHANNEL CONTENTION
- BECAUSE SLOW DEVICES
- BAD PROTOCOLS

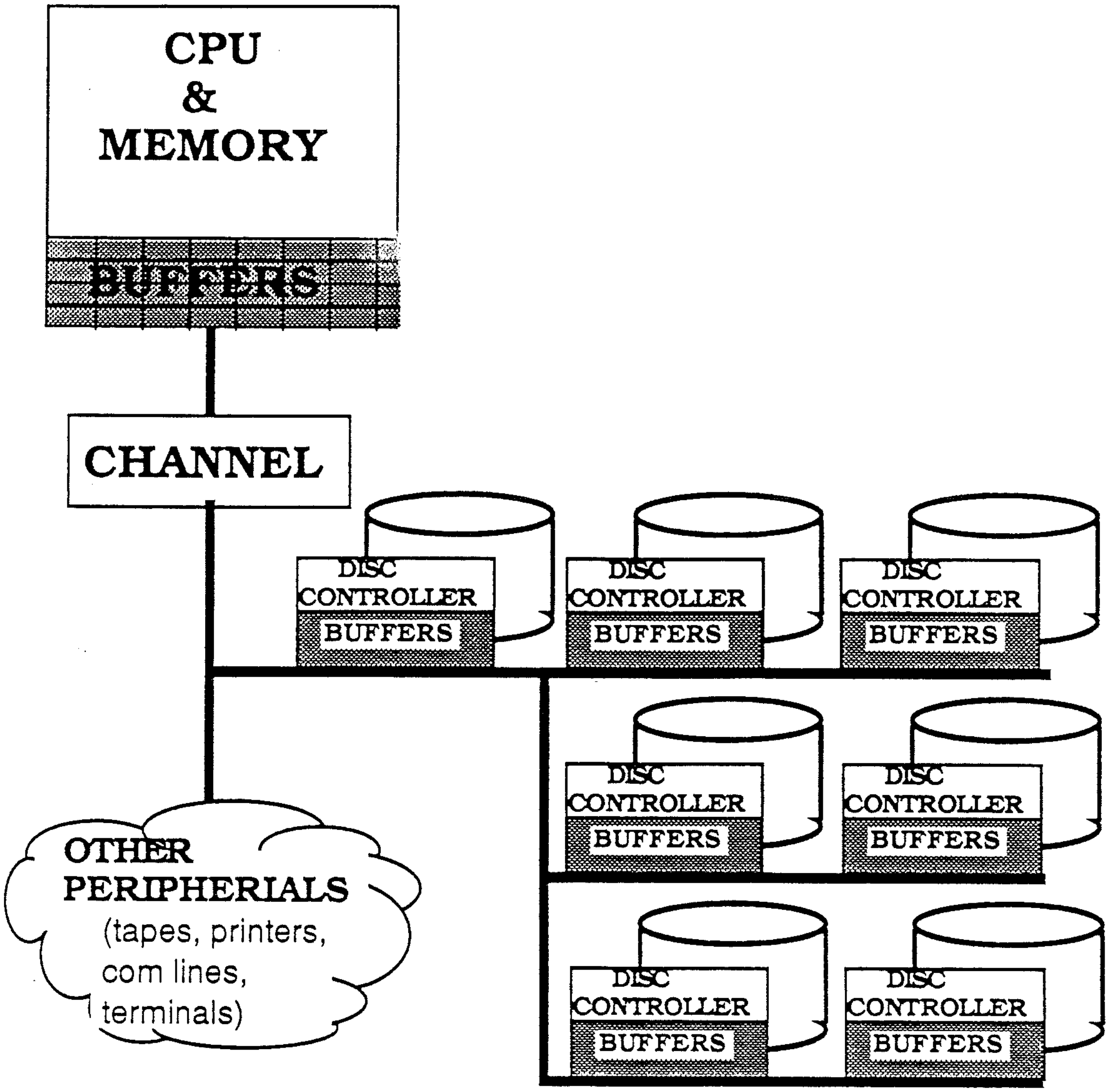
SUGGESTION:
- BUFFER CHANNEL
- BURST MULTIPLEX CHANNEL

HOW TO DESIGN A DISC SUBSYSTEM



TO AVOID QUEUEING WANT MANY
- Arms
- Controllers
- Channels

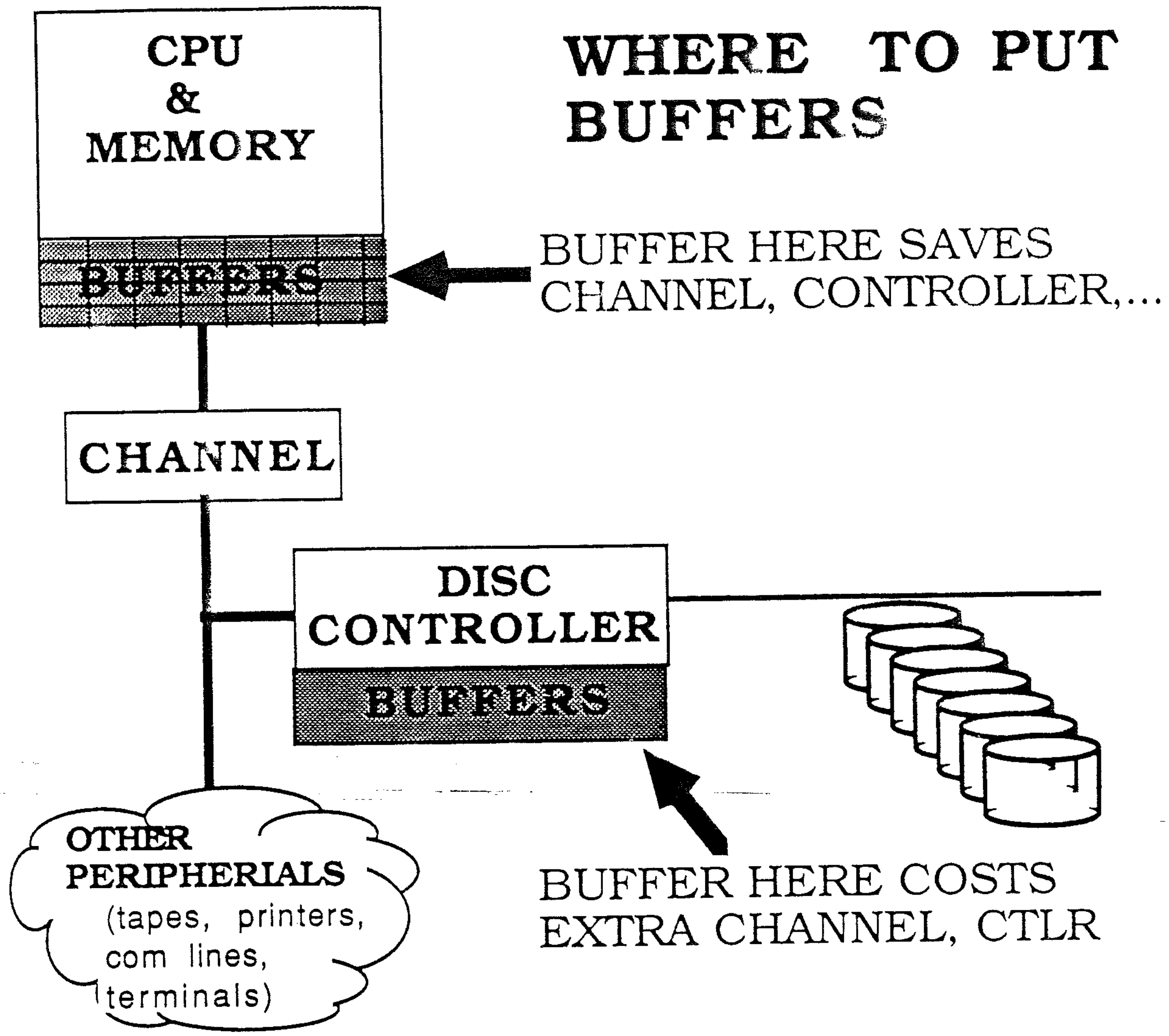
CONTROLLER PER DISC AVOIDS QUEUES



TO AVOID QUEUEING WANT MANY ARMS
CONTROLLERS
CHANNELS

TO AVOID RPS MISS and
TO ALLOW BURST MULTIPLEX CHANNEL WANT
BUFFERED CONTROLLERS

WHERE TO PUT BUFFERS



WHAT IF DISC BUFFER MUCH (10X) CHEAPER

4k PAGE @ 5k$/MB => 20$

4k PAGE @ 500$/MB => 2$

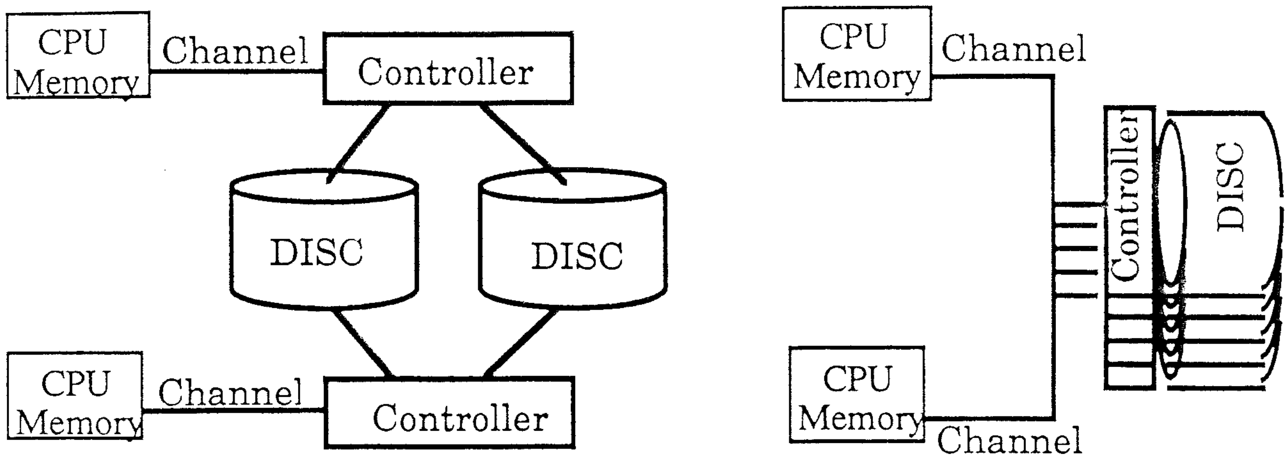
3k ins @ 50K$/MIP => 150$/ACCESS
channel + controller @ 300 a/s
=> 500$/A

BREAK EVEN IS ABOUT 30 SECONDS

SO CASE:

HOT SPOT (RI < 30sec):	MAIN MEMORY
WARM SPOT (RI in [30,1000]):	DISC BUFFER
COLD SPOT (RI > 1000):	DISC

MIRRORED DISCS



- DUAL MODULES (controller, disc)
- DUAL DATA PATHS (4 paths to data)
- READ ANY, WRITE BOTH
- EACH MODULE IS FAIL FAST (disc, controller, path)
- $\text{MTBF}_2 \sim \frac{\text{MTBF}^2}{\text{MTTR}}$

DOES DISC DUPLEXING WORK?

1987 Tandem : 50,000hr MTBF (6 years)
5hr MTTR

=> ~ 65,000 year MTBF

OBSERVED IN LAST 24 MONTHS:
35 double fails on ~46,400 pair/years
~ 1300 years

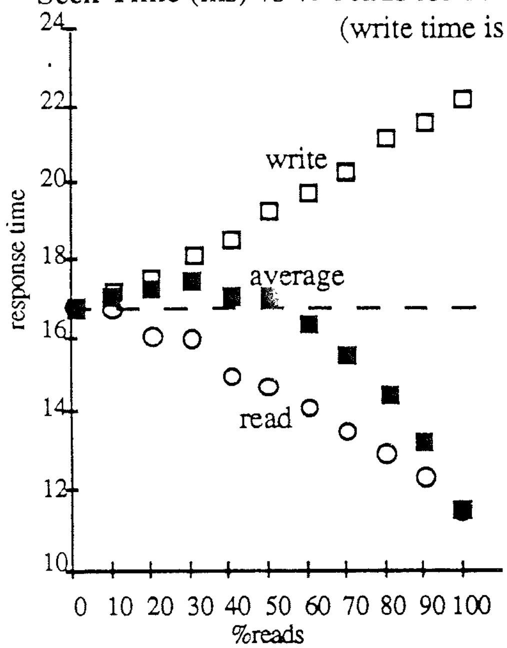
CONCLUSION:

IT WORKS WELL (200x better than no duplex).

FAILURES NOT INDEPENDENT
NOT UNIFORM
INVOLVE CONTROLLERS...
(50x worse than theory)

MIRRORED DISC PERFORMANCE

Seek Time (ms) vs % reads for mirrored discs at low load (no queueing)
(write time is max seek)



The raw data is:

%	read	write	avg
0	16.8	16.8	16.8
10	16.8	17.3	17.2
20	16.0	17.6	17.3
30	15.9	18.1	17.5
40	15.0	18.6	17.1
50	14.7	19.3	17.0
60	14.2	19.8	16.4
70	13.6	20.4	14.6
90	12.4	21.6	13.4
100	11.7	22.2	11.7

Read from closest arm => seek ~$\frac{1}{6}$ tracks

Write farthest arm => seek ~$\frac{1}{2}$ tracks

Mix gives curve above

Note: Shortest service time includes **shortest rotation**

=> save an additional $\frac{1}{6}$ 16 = ~3ms

Total savings on mirrored reads: ~8ms (5+3)

MIRRORED DISC ARM SCHEDULING

Assume FIFO scheduling of requests.

Write scheduling is no-brainer

Read scheduling could be: Shortest Seek

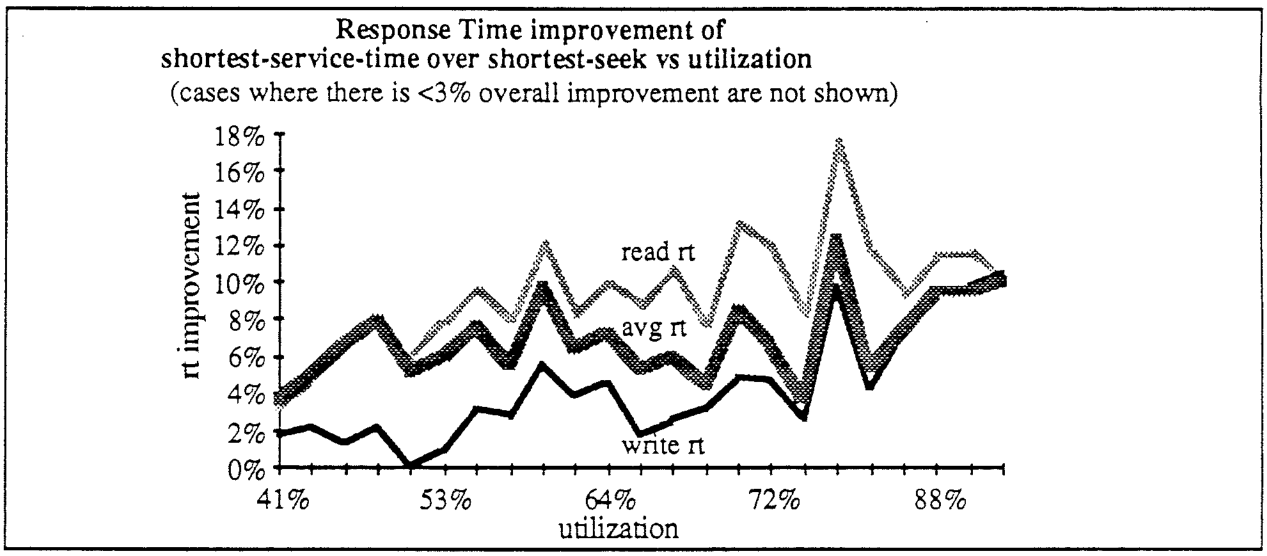
Shortest Service time

Other?

For low loads all are about the same

Between 30% and 90% Shortest Service time is best (~8%)

Read only case:



Even better for mixed reads and writes.

Bitton, D., Gray, J., *Disk Shadowing*, VLDB 1988 Proceedings, Morgan Kauffman, Sept 1988.
Bitton, D., *Arm Scheduling in Shadowed Disks*, COMPCON 1989, IEEE Press, March 1989.
Gray, J., H. Sammer, S. Whitford, Shortest Seek vs Shortest Service Time Scheduling of Mirrored Disc Reads, Tandem Computers December 1988

WHAT ABOUT USING ARRAYS OF SMALL DISCS

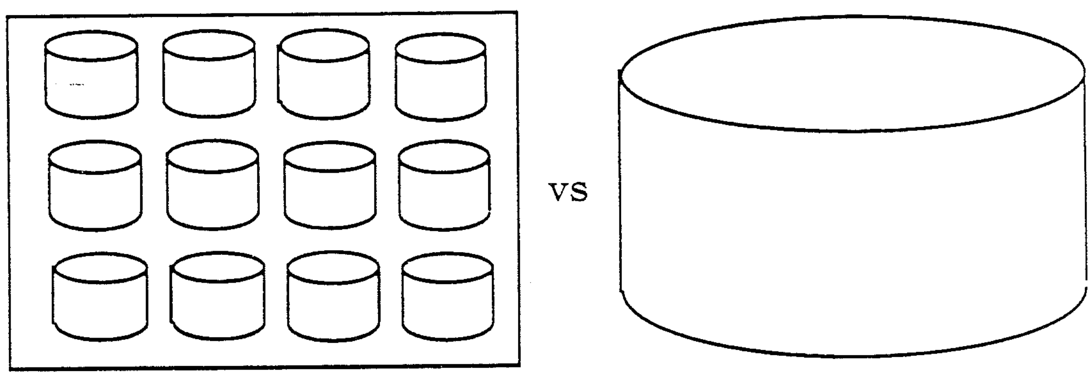
SMALL IS BEAUTIFUL:

MASS PRODUCTION:

LOW COST

DISC IS FIELD REPLACEABLE UNIT

PARALLELISM => performance:
disc striping => 10x bamdwidth



PROBLEM WITH STRIPING:

THE BIG DISC PROBLEM:

Disc Delivers 25accesses/second:

100MB	1 a/s/4MB,
1GB	1 a/s/40MB
10GB	1 a/s/400MB
100GB	1 a/s/4GB

Arms are the scarce/queueing resource
Good if DISC is treated as TAPE: Purely Sequential

WHAT ABOUT USING SMALL DISCS

PROBLEM:

MANY SMALL DISCS => MANY ERRORS

SOLUTIONS:

DUPLEX Discs, Controllers, Paths, Power,...:

Good for small read+writes

RAID (Redundant Arrays of Independent Discs)

N data discs + parity disc.

Good for

Space utilization

read cost (single read if no error)

write cost is 3x (read parity, write data,parity)
compared to duplex 2x

G. Gibson, R. Katz, D. Patterson, *A Case for Redundant Arrays of Inexpensive Discs, (RAID)*, SIGMOD 88.

M. Kim, *Synchronized Discs Interleaving*, IEEE TOC, V. C35 #11, Nov 1986

S. Ng, *Design Alternatives for Disc Duplexing*, IBM RJ 5481, Jan 1987

S. Ng, Lang, D., Sellinger, R., *Tradeoffs Between Devices and Paths In achieving Disc Interleving*, IBM RJ 6140, Mar 1988

S. Ng, *Some Design Issues of Disc Arrays*, Compcon 89

G. Gibson, Peter Chen, R. Katz, D. Patterson, *Introduction To Redundant Arrays of Inexpensive Discs (RAID)*, Compcon 89

M. Schulze, G. Gibson, R. Katz, D. Patterson, *How Reliable is RAID?*, Compcon 89

OUTLINE

DEBIT CREDT STANDARDIZATION

DISC TRENDS & ECONOMICS

DISC PHYSICS

DISC SUBSYSTEMS